1.3.2 Tabla de distribución de frecuencias

Una distribución de frecuencias es una tabla en la cual se agrupan en clases o categorías los valores posibles para unas variables y se registra el número de valores observados que corresponden a cada clase.

Tablas de distribuciones para datos cualitativos

Ejemplo:

Construye una tabla de distribución de frecuencia, para una muestra de compras de refresco según la preferencia de 50 personas

Coke Classic	Sprite	Coke Classic	Pepsi-Cola	Coke Classic	Coke Classic
Pepsi-Cola	Diet Coke	Coke Classic	Diet Coke	Coke Classic	Coke Classic
Coke Classic	Diet Coke	Pepsi-Cola	Coke Classic	Coke Classic	Dr. Pepper
Dr. Pepper	Sprite	Diet Coke	Coke Classic	Diet Coke	Pepsi-Cola
Pepsi-Cola	Coke Classic	Pepsi-Cola	Pepsi-Cola	Coke Classic	Pepsi-Cola
Coke Classic	Coke Classic	Pepsi-Cola	Dr. Pepper	Pepsi-Cola	Pepsi-Cola
Coke Classic	Coke Classic	Coke Classic	Coke Classic	Sprite	Dr. Pepper
Diet Coke	Diet Coke	Pepsi-Cola	Coke Classic	Pepsi-Cola	Sprite
Sprite	Dr. Pepper				

Categoría	Conteo	Frecuencia Real	Frecuencia Relativa	Frecuencia Real Acumulada	Frec. Rel. Acum.

Ejemplo:

Un restaurante de Florida emplea cuestionarios en los que pide a sus clientes que evalúen el servicio, la calidad de los alimentos, los cocteles, los precios y la atmosfera del restaurante. Cada uno de estos puntos se evalúa con una escala de óptimo (O), muy bueno (V), bueno (G), regular (A) y malo (P). Emplee la estadística descriptiva para resumir los datos siguientes respecto a la calidad de los alimentos.

G	O	V	G	A	O	V	O	V	G	O	V	A
V	O	P	V	O	G	A	O	O	O	G	O	V
V	A	G	O	V	P	V	O	O	G	O	O	V
O	G	A	O	V	O	O	G	V	A	G		

Categoría	Conteo	Frecuencia Real	Frecuencia Relativa	Frecuencia Real Acumulada	Frec. Rel. Acum.

Ejercicio 1:
Los seis apellidos más comunes en Estados Unidos, en orden alfabético son, Brown, Davis, Johnson, Jones, Smith y Williams (*The World Almanac, 2006*). Suponga que en una muestra de 50 personas con uno de estos apellidos se obtienen los datos siguientes. Elabore una tabla de distribución

Brown	Williams	Williams	Williams	Brown
Smith	Jones	Smith	Johnson	Smith
Davis	Smith	Brown	Williams	Johnson
Johnson	Smith	Smith	Johnson	Brown
Williams	Davis	Johnson	Williams	Johnson
Williams	Johnson	Jones	Smith	Brown
Johnson	Smith	Smith	Brown	Jones
Jones	Jones	Smith	Smith	Davis
Davis	Jones	Williams	Davis	Smith
Jones	Johnson	Brown	Johnson	Davis

Categoría	Conteo	Frecuencia Real	Frecuencia Relativa	Frecuencia Real Acumulada	Frecuencia Relativa Acumulada

Ejercicio 2:
Los cuatro programas con horario estelar de televisión son CSI, ER, Everybody Loves Raymond y Friends (*Nielsen Media Research,* 11 de enero de 2004). A continuación se presentan los datos sobre las preferencias de los 50 televidentes de una muestra. Elabore una tabla de distribución

CSI	Friends	CSI	CSI	CSI
CSI	CSI	Raymond	ER	ER
Friends	CSI	ER	Friends	CSI
ER	ER	Friends	CSI	Raymond
CSI	Friends	CSI	CSI	Friends
ER	ER	ER	Friends	Raymond
CSI	Friends	Friends	CSI	Raymond
Friends	Friends	Raymond	Friends	CSI
Raymond	Friends	ER	Friends	CSI
CSI	ER	CSI	Friends	ER

Categoría	Conteo	Frecuencia Real	Frecuencia Relativa	Frecuencia Real	Frecuencia Relativa

				Acumulada	**Acumulada**

Elaboración de una tabla de distribución de frecuencias

Una distribución de frecuencias del tipo "y menor que" se construye de la siguiente manera:

1. Se determina la amplitud de variación de datos, detectando el valor más grande y el más pequeño del conjunto y calculando su diferencia.

$$Av = X_{max} - X_{min}$$
$$X_{max} = Es\ el\ valor\ mas\ grande\ del\ conjunto$$
$$X_{min} = Es\ el\ valor\ mas\ pequeño\ del\ conjunto$$

2. Determinar el número de clases o categorías a utilizar.

$$K = \sqrt{n}\ Redondeando\ al\ entero\ mas\ cercano, donde:$$

n= Es el tamaño de muestra

3. Determinar el ancho de cada clase o categoría:

$$C = \frac{Av}{K}$$

Redondeando al entero más grande.

4. Establecer los límites de cada categoría. El límite inferior de la primera categoría (LI_1) debe ser un poco más pequeño que el valor mínimo. El límite superior (LS_1) de esta categoría se calcula de la siguiente manera:

$$LS_1 = LI_1 + C$$

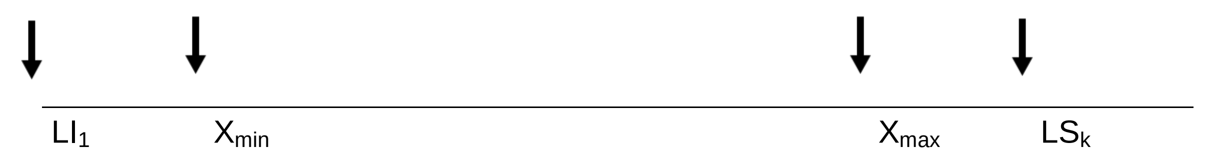
El límite inferior de la segunda categoría (LI_2) será el LS_1 y el límite superior de esta segunda categoría será:

$$LS_2 = LI_2 + C$$

De manera general, entonces el límite superior de la i-esima categoría seria:

$$LS_i = LI_i + C \qquad \therefore \qquad i = 1,2,\ldots,k$$

Es importante que los valores máximos y mínimo de la muestra caigan dentro de los límites inferior y superior de la primera y última categoría respectivamente, y que además estén centrados, es decir, que haya la misma distancia del LI_1 al X_{max}, por ejemplo:



Puesto que los límites se calculan utilizando un procedimiento de ensayo y error, se sugiere utilizar la siguiente fórmula para calcular el límite superior de la última

categoría, y así determinar cual sería la mejor combinación de valores de ancho de clase, y límite inferior:

$$LS_k = KC + LI_1$$

5. Una vez que están centrados los limites, se empiezan a registrar los datos en la categoría que corresponden, haciendo una marca por cada dato en la categoría que le corresponde, tomando el criterio de registrar un dato en la categoría 1 por ejemplo si este es < LS_1, o registrar un dato en la categoría 2 si es <LS_2 y así sucesivamente.

6. Construir la tabla de distribución de frecuencias que contenga:
   - ✓ Numero de clases
   - ✓ Categoría o clase
   - ✓ Punto medio
   - ✓ Conteo
   - ✓ Frecuencia real
   - ✓ Frecuencia relativa

Ejemplo 16:

Considere la resistencia a la tensión de 80 muestras de aleación de aluminio-litio.

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Núm. De Clase	Intervalo de Clase	Punto Medio	Conteo	Frec. Real	Frec. Relativa
TOTAL DE OBSERVACIONES					

Ejercicio 2:

Las siguientes son medidas de las resistencias de la resistencia rompimiento (en onzas) de una muestra de 60 hilos de lino

32.5	15.2	35.4	21.3	28.4	26.9	34.6	29.3	24.5	31.0
21.2	28.3	27.1	25.0	32.7	29.5	30.2	23.9	23.0	26.4
27.3	33.7	29.4	21.9	29.3	17.3	29.0	36.8	29.2	23.5
20.6	29.5	21.8	37.5	33.5	29.6	26.8	28.7	34.8	18.6
25.4	34.1	27.5	29.6	22.2	22.7	31.3	33.2	37.0	28.3
36.9	24.6	28.9	24.8	28.1	25.4	34.5	23.6	38.4	24.0

Núm. De Clase	**Intervalo de Clase**	**Punto Medio**	**Conteo**	**Frec. Real**	**Frec. Relativa**
TOTAL DE OBSERVACIONES					

1.3.3 Medidas de Tendencia Central y de Dispersión

Hay que recordar que todas las medidas de tendencia central y de dispersión obtenidas a partir de una tabla de distribución de frecuencias, no son medidas exactas, ya que se generan a partir de un conjunto de datos ya procesados, por lo que tienen una ligera pérdida de información con respecto a las medidas obtenidas directamente a partir del grupo de datos sin procesar, en el cual se toman en cuenta todos los datos del conjunto para obtener estas medidas.

Media

Esta medida de tendencia central basada en la distribución de frecuencias se obtiene aplicando la siguiente fórmula de cálculo:

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i \times P.M}{n}$$

$$f_i = Frecuencia\ real\ de\ la\ i-ésima\ clase\ o\ categoría$$

$$P \cdot M_i = punto\ medio\ de\ la\ i-ésima\ categoría$$

$$k = número\ de\ categorías$$

$$n = total\ de\ datos\ en\ la\ muestra$$

Ejemplo de Resistencia al Rompimiento:

No. Clase	Punto Medio	Frecuencia Real	$(\boldsymbol{P.M.})(\boldsymbol{fi})$
1	16.7	2	
2	19.7	2	
3	22.7	11	
4	25.7	10	

5	28.7	17	
6	31.7	5	
7	34.7	8	
8	37.7	5	
$\sum =$		60	

Mediana

Esta medida divide en dos partes iguales al conjunto de datos, sin embargo al obtenerla a partir de la tabla de distribución de frecuencias, la exactitud en la división se pierde, y de manera aproximada se calcula la división. Para obtener esta medida es necesario primero encontrar la clase mediana; esta es aquella cuya frecuencia acumulada real es igual o mayor que la mitad del conjunto de datos. Una vez identificada esta clase mediana se aplica la siguiente fórmula de cálculo:

$$\tilde{X} = LI + \left[\frac{\left(\frac{n}{2}\right) - f_{a_A}}{f}\right] c$$

$$LI = \text{límite inferior de la clase mediana}$$

$$f_{a_A} = \text{Frecuencia real acumulada anterior a la clase mediana}$$

$$f = \text{Frecuencia real de la clase mediana}$$

$$c = \text{amplitud de la clase}$$

Ejemplo de Resistencia al Rompimiento:

No. Clase	Intervalo de Clase	Frecuencia Real	Frecuencia Relativa	Frecuencia Real Acumulada	Frecuencia Relativa Acumulada
1	15.2-18.2	2	0.0333	2	0.0333
2	18.2-21.2	2	0.0333	4	0.0666
3	21.2-24.2	11	0.1833	15	0.2499
4	24.2-27.2	10	0.1666	25	0.4165
5	27.2-30.2	17	0.2833	42	0.6998
6	30.2-33.2	5	0.0833	47	0.7831
7	33.2-36.2	8	0.1333	55	0.9164
8	36.2-39.2	5	0.0833	60	0.9997

Nota: la clase mediana es aquella cuya frecuencia real acumulada iguala o excede a la mitad del total de las observaciones.

Moda por Interpolación

$$L\hat{X} = LI + \left[\frac{d_1}{d_1 + d_2}\right] c$$

$$L\hat{X} = \text{Moda por interpolación}$$

$$LI = \text{Límite inferior de la clase modal (la clase de mayor frecuencia)}$$

$$d_1 = \text{diferencia entre la clase modal y la clase anterior a la modal}$$

$$d_2 = \text{diferencia entre la clase modal y la clase posterior a la modal}$$

$$c = amplitud\ de\ clase$$

Ejemplo de Resistencia al Rompimiento:

1.3.3 Medidas de Dispersión

Las medidas de dispersión más comunes en un conjunto de datos a partir de una tabla de distribución de frecuencias son:

Varianza (s^2) y Desviación estándar (s)

A partir de una tabla de distribución de frecuencias y considerando que esta fue construida a partir de una muestra, la varianza y la desviación estándar se obtienen respectivamente como sigue:

$$s^2 = \frac{\sum_{i=1}^{k}(P.M.i)^2 \times f_i - \frac{\left(\sum_{i=1}^{k} P.M.i \times f_i\right)^2}{n}}{n-1}$$

$$s = \sqrt{\frac{\sum_{i=1}^{k}(P.M_i)^2 \times f_i - \frac{\left(\sum_{i=1}^{k} P.M_i \times f_i\right)^2}{n}}{n-1}}$$

$$P.M_i = Punto\ Medio\ de\ la\ i - \acute{e}sima\ categoría$$
$$f_i = Frecuencia\ real\ de\ la\ i - \acute{e}sima\ categoría$$
$$n = tamaño\ de\ muestra$$

Varianza
Ejemplo de la Resistencia al Rompimiento:

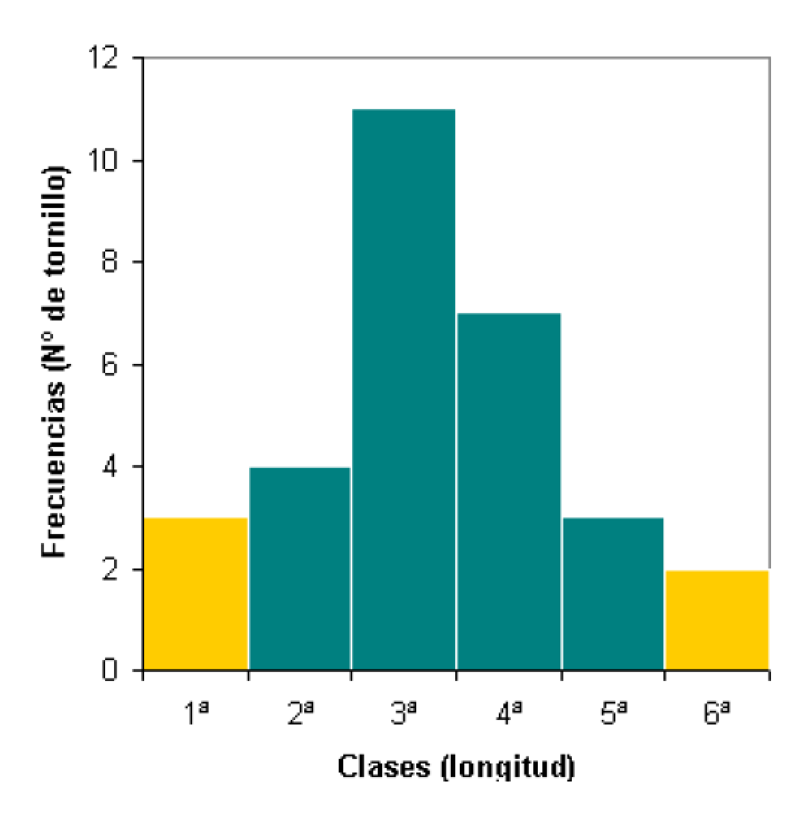
No. Clase	Punto Medio	Frecuencia Real	$(P.M.)(fi)$	$(P.M.)^2(fi)$
1	16.7	2		
2	19.7	2		
3	22.7	11		
4	25.7	10		
5	28.7	17		
6	31.7	5		
7	34.7	8		
8	37.7	5		
$\sum =$		60		

Desviación Estándar
Ejemplo de la Resistencia al Rompimiento:

1.4 Representaciones Graficas

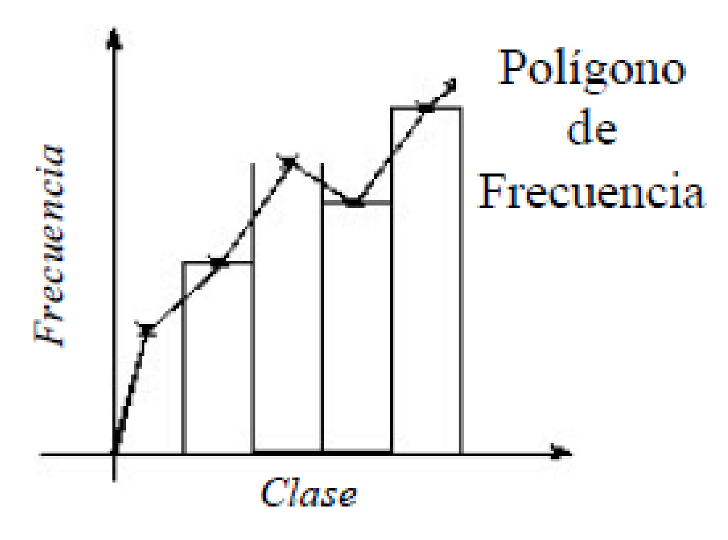
Histogramas

Un histograma es la grafica de barras de una distribución de frecuencias, y en su construcción se colocan sobre el eje horizontal de la grafica los limites exactos de clase, en tanto que sobre el eje vertical se coloca el numero o porcentaje de observaciones correspondientes a cada clase.



Polígono de Frecuencias

Un polígono de frecuencias es la grafica lineal de una distribución de frecuencias, los dos ejes de esta grafica son similares a los del histograma, excepto que, se coloca el punto medio de cada clase sobre el eje horizontal. El numero de observaciones en cada clase se representa por un punto en el punto medio de la clase y estos puntos están unidos por una serie de segmentos de línea para formar una “figura de varios lados” o polígono.



Problema 1:

El comisionado de transporte del condado de Orange está preocupado por la velocidad a la que los conductores manejan en un tramo de la carretera principal. Los datos de velocidad de 45 conductores son los siguientes:

15	32	45	46	42	39	68	47	18
31	48	49	56	52	39	48	69	61
44	42	38	52	55	58	62	58	48
56	58	48	47	52	37	64	29	55
38	29	62	49	69	18	61	55	49

a) Elabore la tabla de distribución de frecuencias
b) Calcule: La media, mediana, moda, varianza y desviación estándar para datos agrupados
c) Elabore el histograma de frecuencias
d) Elabore el polígono de frecuencias

Problema 2:

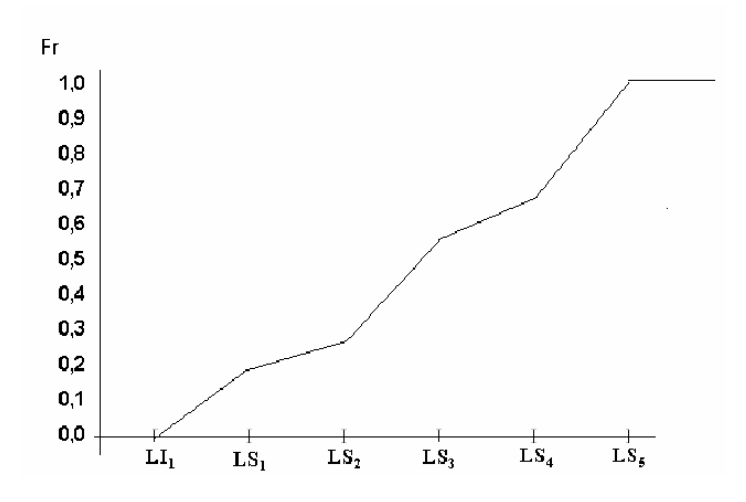
Nielsen Home Technology Report informa sobre la tecnología en el hogar y su uso. Los datos siguientes son las horas de uso de computadora por semana en una muestra de 50 personas

4.1	1.5	10.4	5.9	3.4	5.7	1.6	6.1	3.0	3.7
3.1	4.8	2.0	14.8	5.4	4.2	3.9	4.1	11.1	3.5
4.1	4.1	8.8	5.6	4.3	3.3	7.1	10.3	6.2	7.6
10.8	2.8	9.5	12.9	12.1	0.7	4.0	9.2	4.4	5.7
7.2	6.1	5.7	5.9	4.7	3.9	3.7	3.1	6.1	3.1

a) Elabore la tabla de distribución de frecuencias
b) Calcule: La media, mediana, moda, varianza y desviación estándar para datos agrupados
c) Elabore el histograma de frecuencias
d) Elabore el polígono de frecuencias

Ojiva

También llamado polígono de frecuencias acumuladas, se emplea en distribuciones de frecuencias cuyas clases son intervalos. Es un tipo especial de gráfico de curvas en el cual se representan las frecuencias acumuladas. En el eje horizontal se marca sucesivamente los límites superiores de cada clase y en el vertical las frecuencias acumuladas. Para cada límite superior de clase se marca con un punto su correspondiente frecuencia acumulada y al límite inferior de la primera clase se le asigna una frecuencia acumulada igual a 0. Se unen todos los puntos con segmentos de recta y se obtiene una curva no decreciente.



Grafica Pastel

Una grafica de pastel es especialmente apropiada para ilustrar las divisiones de una cantidad total, por ejemplo la distribución de ingresos o egresos en una compañía. Para construir este tipo de grafica es necesario multiplicar la frecuencia relativa de cada categoría por 360°, para determinar el tamaño de cada sección.

Ejemplo 18:

La siguiente tabla se basa en los datos publicados en los Indicadores Económicos del Banco de México. Y son datos preliminares correspondientes a Diciembre de 1988.

CATEGORIA	**CANTIDAD**	**FREC. RELATIVA**	**TAMAÑO DE SECTOR EN GRADOS**
Petrolera	$560.1	0.33	
Agropecuarias	$143.0	0.08	
Extractivas	$49.9	0.03	
Manufactureras	$951.1	0.56	
Total	$1704.1	1	

1.5 Otras Graficas

Diagrama de Puntos

Un diagrama de puntos se parece a un histograma en que representa de manera grafica la distribución de los valores de los datos; la diferencia es que los valores se grafican de forma individual, en vez de agruparse por categorías o clases. Estos diagramas son útiles para representar conjuntos pequeños de datos que no requieren ser agrupados en categorías; además son especialmente útiles en la comparación de dos conjuntos de datos o dos subgrupos de un mismo conjunto de datos.

En este ejemplo, cada punto representa el peso de una persona y la altura de la misma persona.

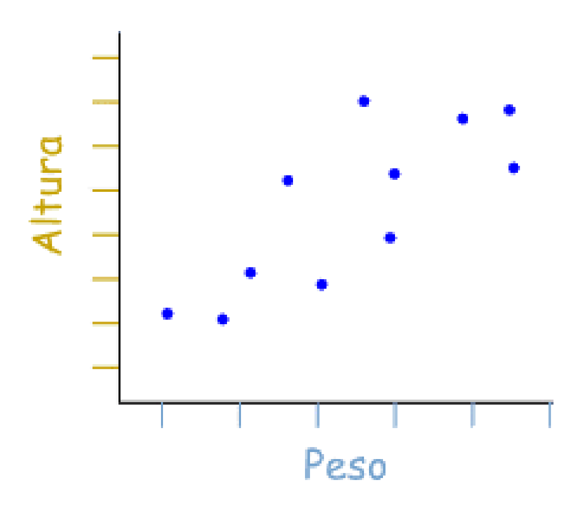


Diagrama de Caja

La grafica de caja es muy útil para reflejar propiedades de una muestra, ofrece una representación visual del conjunto de datos basada en el resumen de cinco números: el valor mínimo del conjunto (X_{min}), el cuartil 1 inferior (Q_1), la mediana del conjunto de datos, el cuartil superior (Q_3) y el valor máximo del conjunto (X_{max}). Construida de esta manera la grafica de caja; contiene el 50% de los datos dentro de la caja considerando el rango intercuartílico, el 25% de los datos fuera de la caja con respecto al valor mínimo (X_{min}) y el otro 25% con respecto al valor máximo (X_{max}). Estos porcentajes de datos fuera de la caja se representan a través de una línea recta a cada lado denominada bigote o extensión. La longitud de los bigotes indicaría si la distribución de los datos es asimétrica o no; y si lo es, nos diría también el tipo de asimetría.

Ejemplo 19:
Se tienen los siguientes datos:
36, 25, 37, 24, 39, 20, 36, 45, 31, 31, 39, 24, 29, 23, 41, 40, 33, 24, 34, 40
Elabora el diagrama de caja

El bigote de la izquierda representa (X_{min}, Q_1)
La primera parte de la caja (Q_1, Q_2)
La segunda parte de la caja (Q_2, Q_3)
El bigote de la derecha (Q_3, X_{max})